sPPM and UMT2K on BGL

Gyan Bhanot
IBM Research

LLNL Applications

SPPM – 3d compressible hydrodynamics
Problem size per node fixed, Grid Comm.
Tests weak scaling

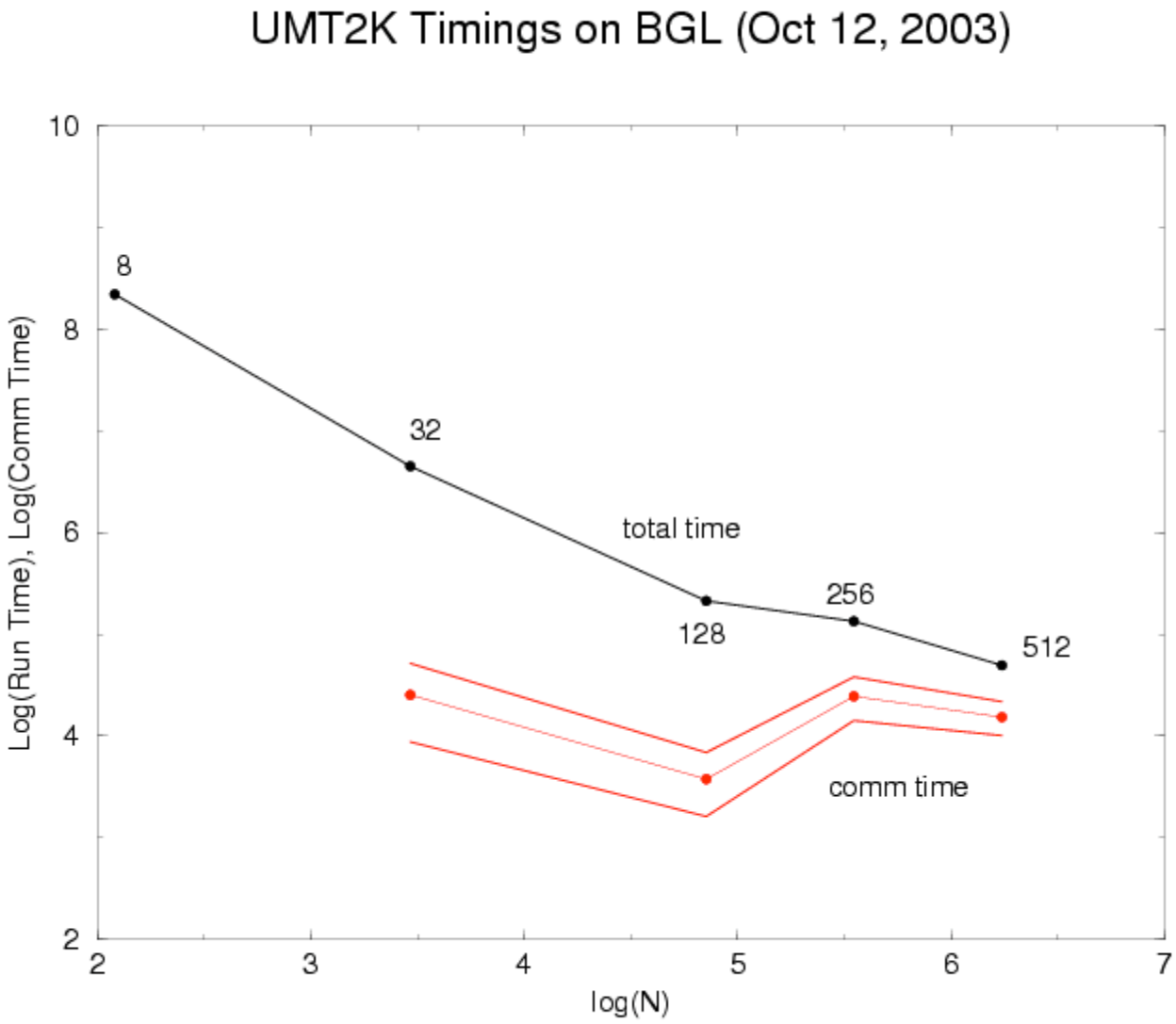
UMT2K - unstructured mesh radiation transport
Problem size fixed, domain decomposition,
Node communicates to many neighbors.
Tests strong scaling

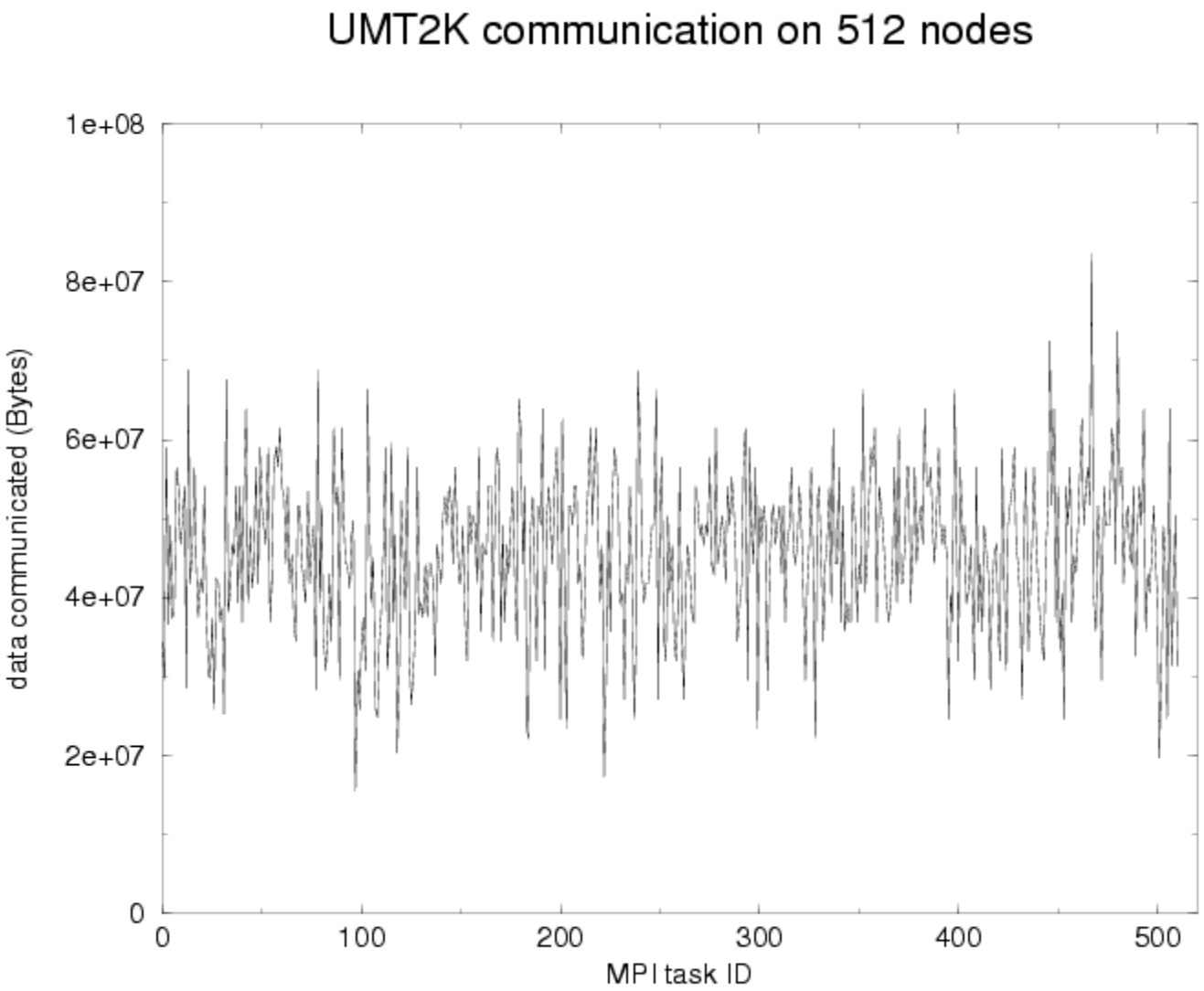
Ran Optimized routines using MPI and blrts_xlc, blrts_xlf

Tested on 8, 32,128, 256, 512 way

UMT2K Timings (10-12-03)

Nodes	Elapsed (seconds)	Commun (seconds)	Compute (seconds)
512	109	66 +/- 11	43 +/- 11
256	168	80 +/- 17	88 +/- 17
128	206	36 +/- 11	170 +/- 11
32	772	82 +/- 30	691 +/- 30
8	4200		


UMT2K Timings on BGL (Oct 12, 2003)
Log(Run Time), Log(Comm Time)
log(N)
10
8
6
4
2
2
3
4
5
6
7
8
32
128
256
512
total time
comm time

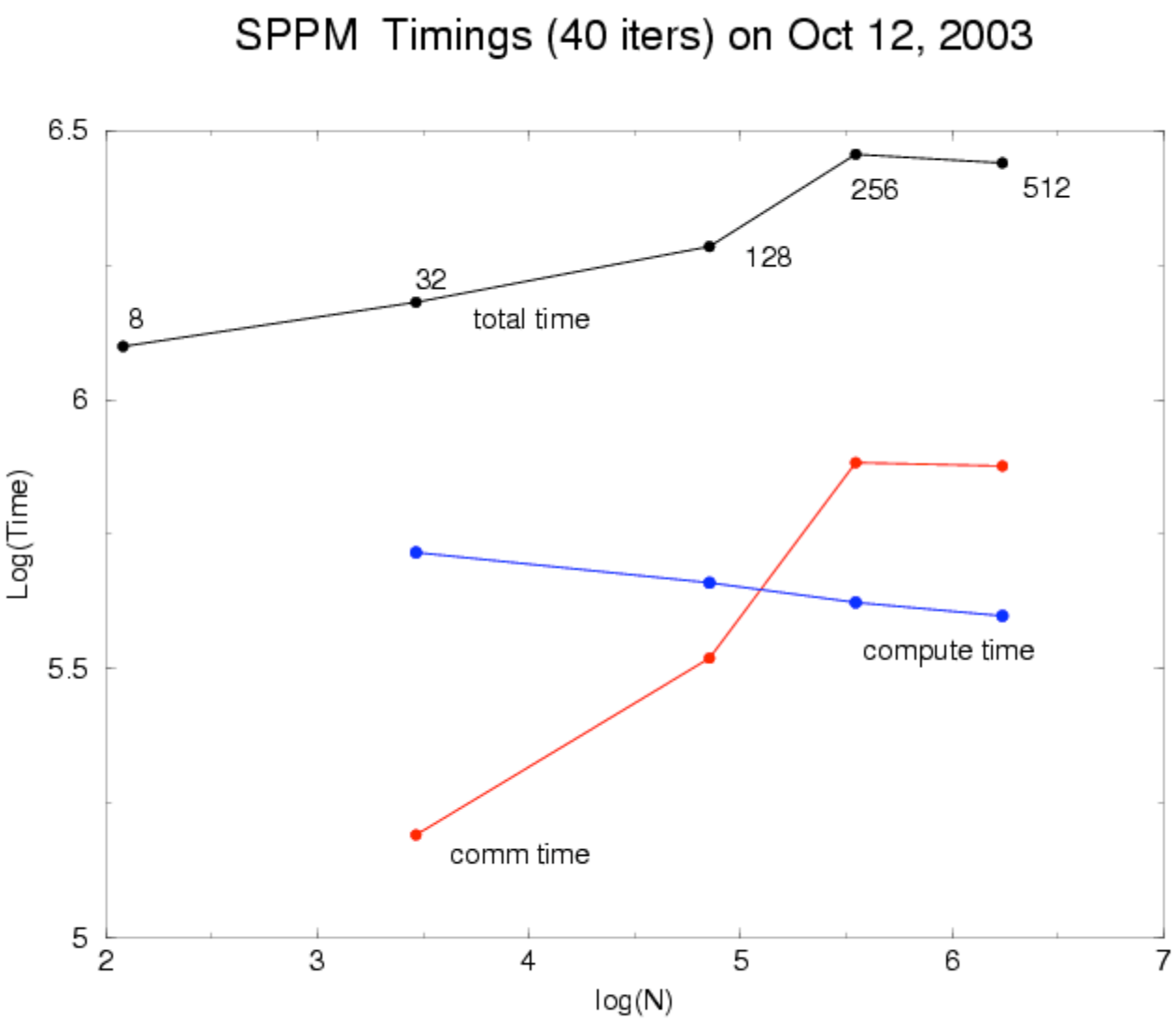

UMT2K communication on 512 nodes
data communicated (Bytes)
1e+08
8e+07
6e+07
4e+07
2e+07
0
0
100
200
300
400
500
MPI task ID

Whats Going on in UMT2K?

- For > 512 nodes test problem is too small. There is not enough work to do per node.
- Problem made worse by load Imbalance. Nodes communicate different amounts of data to different numbers of nodes
- May improve using improved MPI task remapping techniques.

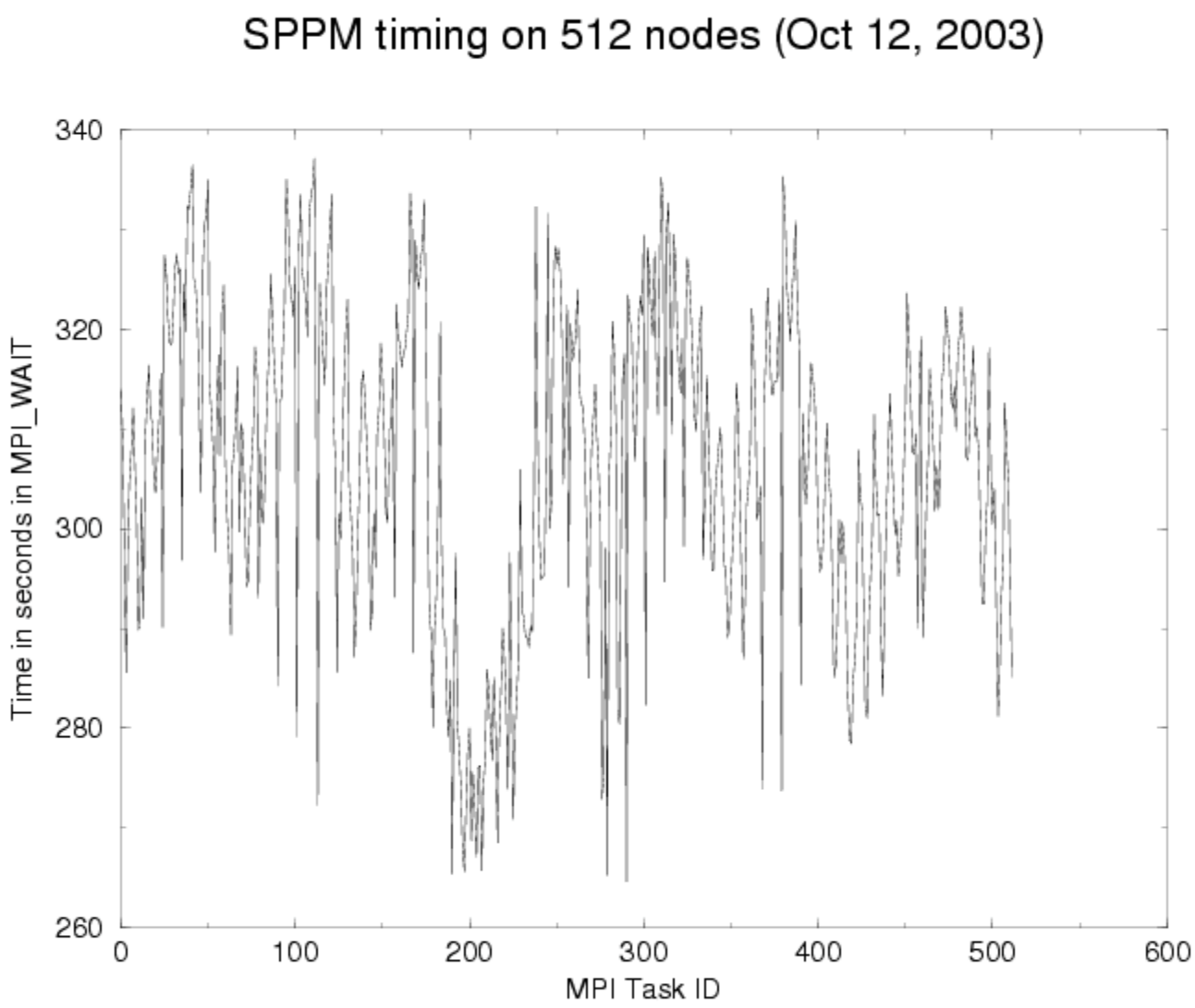
sPPM timings (10-12-03)

Nodes	Elapsed (seconds)	Commun. (seconds)	Compute (seconds)
512	626	357 +/- 15	269 +/- 15
256	636	359 +/- 16	277 +/- 16
128	537	250 +/- 17	287 +/- 17
32	484	180 +/- 11	304 +/- 11
8	445		


SPPM Timings (40 iters) on Oct 12, 2003
6.5
6
5.5
5
2
3
4
5
6
7
Log(Time)
log(N)
8
32
128
256
512
total time
compute time
comm time

What is Going On in sPPM?

- Currently, system uses only single CPU for compute and communication
- MPI_ISEND and MPI_IRECV just return
- All communication is done at MPI_WAIT
- sPPM is written to do computation after MPI_IRECV, MPI_ISEND are posted to overlap computation and communication
- Results in lots of time at MPI_WAIT


SPPM timing on 512 nodes (Oct 12, 2003)
Time in seconds in MPI_WAIT
340
320
300
280
260
0
100
200
300
400
500
600
MPI Task ID